国立国会図書館デジタル化資料データ（画像・テキスト）の使い道

90分アイデアソンの進め方

16時～16時15分　アイデアソン説明【15分】

アイデアソンのインプットとして、当館のOCR関連事業、オープンデータ及び実験サービスを紹介

16時15分～17時　ブレイクアウトセッション【45分】

【アイデアソン参加者】

3班×5-6名に分かれ、各ブレイクアウトルームで議論を行う

【一般参加者（聴講者）】

- アイデアソン参加者の議論の様子は、オンラインホワイトボードの共有により確認できる
- モデレータが最初に導入的なプレゼンを行い、次にホワイトボードを見ながら、各ブレイクアウトルームの議論の様子を視聴者向けに紹介＆コメントをする

17時～17時30分　アイデアソン参加者各班の発表、総括【30分】

各班から議論の結果を発表し（発表4分＋質疑4分）、モデレータが講評及び全体の総括を行う

NDL Lab
National Diet Library, Japan

国立国会図書館デジタル化資料データ（画像・テキスト）の使い道：90分アイデアソン インプット資料

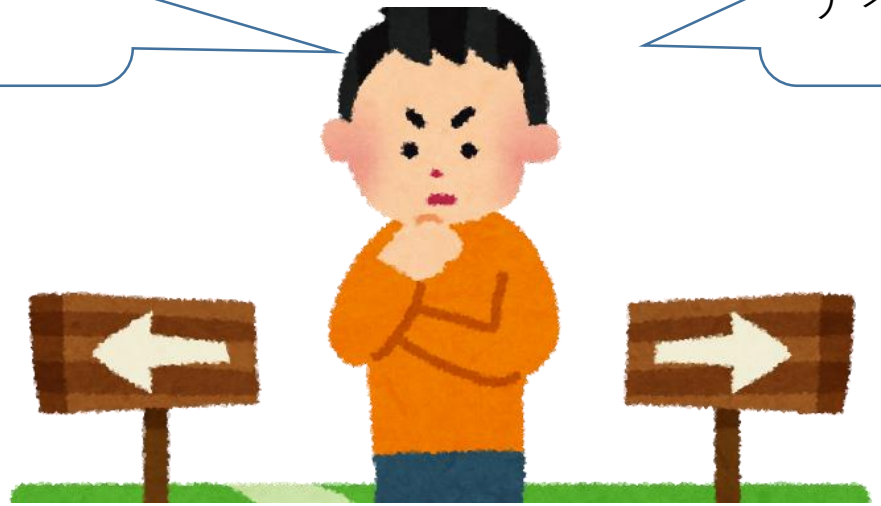
国立国会図書館電子情報部電子情報企画課
次世代システム開発研究室　青池亨

NDL Lab
National Diet Library, Japan

この資料の見方

データを検索しながらチラ見して
アイデアを膨らませたい

まとまったデータをダウンロードして手元で
アイデアを膨らませたい



【サービス】とついている
スライドをご覧ください

【データセット】とついている
スライドをご覧ください
（※データセット①を除く）

本日のアイデアソンのために詳しく説明するデータセット

国立国会図書館（NDL）では、令和2年度補正予算（第3号）により、令和3年度に2つの事業を実施しました

（参考:https://lab.ndl.go.jp/data_set/ocr/）

事業1. 当館がこれまで提供していたほぼ全ての**デジタル化資料（約247万点）を**当館資料用に特別に性能改善した**OCRでテキスト化**する事業

事業2. 今後当館がデジタル化する資料に対してテキスト化処理を行う**OCR処理プログラムを研究開発**する事業

これらの事業によって、新たに提供可能になったデータセットを主に取り上げます

NDLが提供しているその他のオープンデータセット（参考情報）

「書誌データ」「典拠データ」
「国会会議録テキストデータ・画像データ」
「全国の図書館や公文書館によるレファレンス事例」
といったものがあります

- https://www.ndl.go.jp/jp/dlib/standards/opendataset/index.html
- https://lab.ndl.go.jp/data_set/dataset/

本日は、OCR事業の成果物であるデータセットを中心にお話ししますが、是非これらのデータ資源もご活用ください！

【データセット①】
OCRテキストデータ

どんなデータセット？

- 事業1で作成された、247万点分の全文テキストデータ
- 座標情報や縦書き・横書きといった情報も含まれる
- 自動で読み取られた結果なので、一定の割合で誤りがある

データセットのサンプル

https://lab.ndl.go.jp/dataset/joss2022/OCRtxtdatasample.zip

本日の「主菜」
とはいえ、分量が膨大なので、生データをぽんと渡されても、一日では把握しきれないと思います
中身をつまみ食いできる実験サービスを後ほどご紹介します

【データセット①】
OCRテキストデータ

OCRテキストデータが持つ情報（テキストボックス単位の情報）

Key	Description
words	テキストボックス情報のリスト
id	テキストボックスの順序
boundingBox	テキストボックスの座標位置
isVertical	縦書きか横書きか
text	文字列単位に区切られたテキストデータ 縦書きの場合は文字が正立する向きで上から、横書きの場合は左から順に出力される
confidence	文字認識の信頼度 0-1の範囲を取り、大きいほど信頼度が高い
isTextline	本文か、それ以外か（文字サイズによる判定）
isRTL	右横書きか、左横書きか

【データセット①】
OCRテキストデータ

OCRテキストデータが持つ情報（1行単位の情報）

Key	Description
lines	行ボックス情報のリスト
id	行ボックスの順序
boundingBox	行ボックスの座標位置
wordIDs	行に含まれるテキストボックスのidリスト
estimatedLanguage	言語の推定結果 ja: 日本語, ko: 韓国語, ta: 台湾語, en: 英語 * テキストボックスが1つも検出されなかった場合、“NULL”(文字列)を出力

【データセット②】
OCR処理プログラム（NDLOCR）

※便宜上データセットと書きましたが、プログラムです

https://github.com/ndl-lab/ndlocr_cli

どんなプログラム？

- デジタル化資料に記述された文字列を読み取ってテキスト化するプログラム
- そのまま使うにはシステムの知識が必要だが、東京大学の中村覚助教らにより、簡単に使えるチュートリアルが用意されている

https://zenn.dev/nakamura196/articles/b6712981af3384

【データセット③】
OCR学習用データセット

どんな人に向いている？

自分で機械学習を学んでOCR処理プログラムを作りたい方や、
今利用しているOCRの性能を評価したい方

どんなデータセット？

- 資料画像
- 資料画像の内部に書かれた正解テキスト情報
- 一部のデータにはレイアウト情報（キャプション、タイトル、著者名等）も入っている

【データセット③】
OCR学習用データセット

- https://github.com/ndl-lab/pdmocrdataset-part1

OCRテキスト化事業の性能改善を目的として、当館の保有するデジタル化資料から作成したOCR学習用途の機械学習データセットのうち、 著作権保護期間の満了した資料から作成されたデータセット（2,713画像分）

- https://github.com/ndl-lab/pdmocrdataset-part2

NDLOCRの学習を目的として当館の提供するデジタル化資料から作成したOCR学習用途の機械学習データセットのうち、著作権保護期間の満了した資料から作成されたデータセット（3,997画像分）

【データセット④】
レイアウトデータセット・図版タグデータセット

著作権保護期間満了資料から、NDLラボが作成した画像のデータセット

NDL-DocL（資料画像レイアウトデータセット）

https://github.com/ndl-lab/layout-dataset

→古典籍資料と明治以降刊行資料についてそれぞれ作成

NDL-ImageLabel（ラベル付画像データセット）

https://github.com/ndl-lab/imagetagdataset

→自動で切り出された図版を、写真の種類やイラストの種類で分類してタグ付けしたデータセット

【サービス①】
次世代デジタルライブラリー

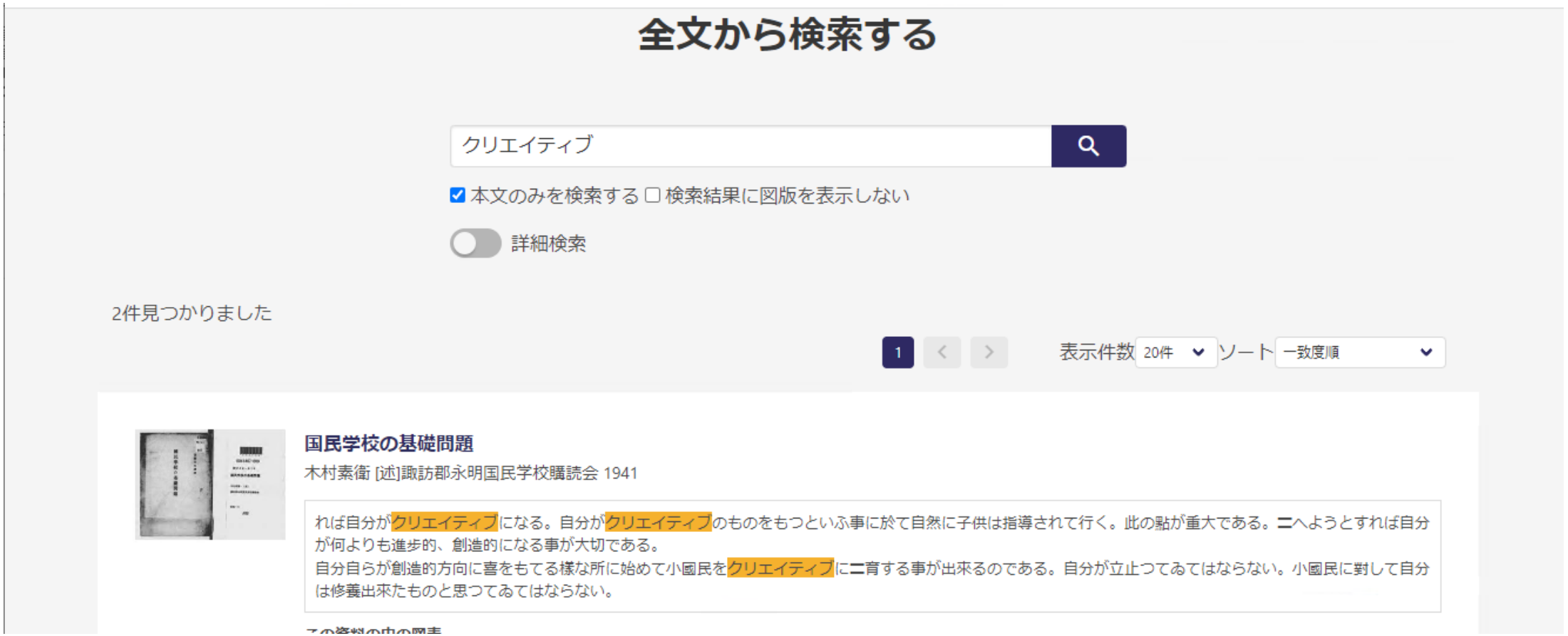
https://lab.ndl.go.jp/dl/

国立国会図書館デジタルコレクションで提供しているデジタル化資料の中から、著作権の保護期間が満了した図書及び古典籍資料全部（約33万6千点）が検索可能（全文検索は図書のみ約28万点）な実験サービス

何ができるサービス？

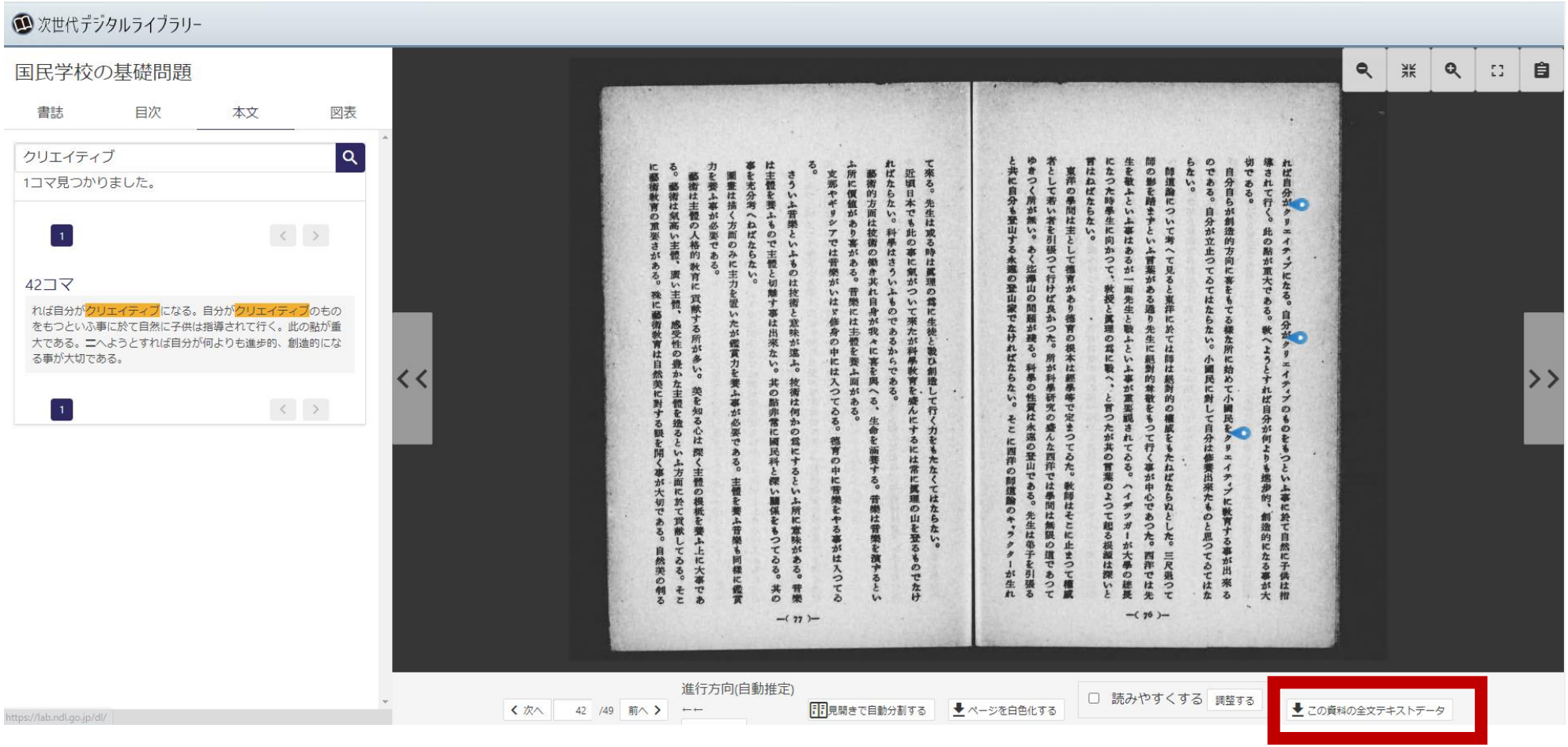
- **デジタル化資料の「中身」を全文検索&画像検索できる**
- **資料ごとにOCRテキストデータをダウンロードできる**

【サービス①】
次世代デジタルライブラリー



https://lab.ndl.go.jp/dl/fulltext?keyword=クリエイティブ&searchfield=contentonly

【サービス①】
次世代デジタルライブラリー

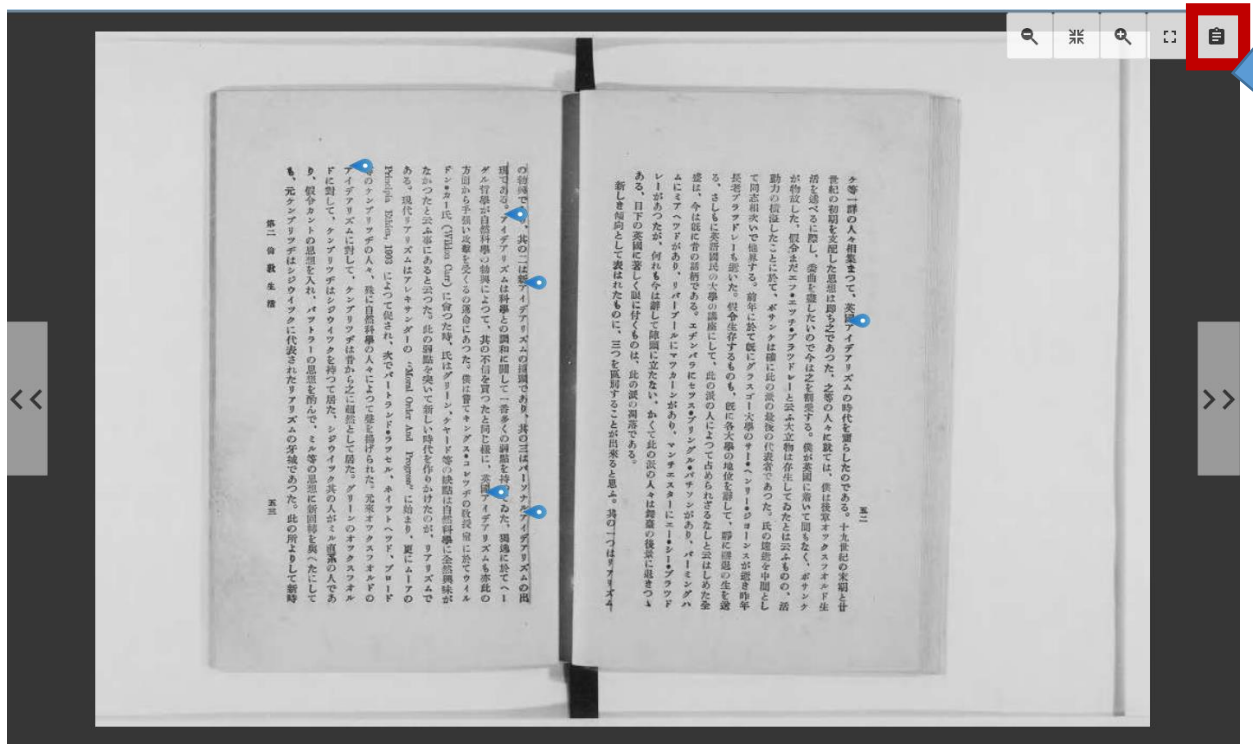


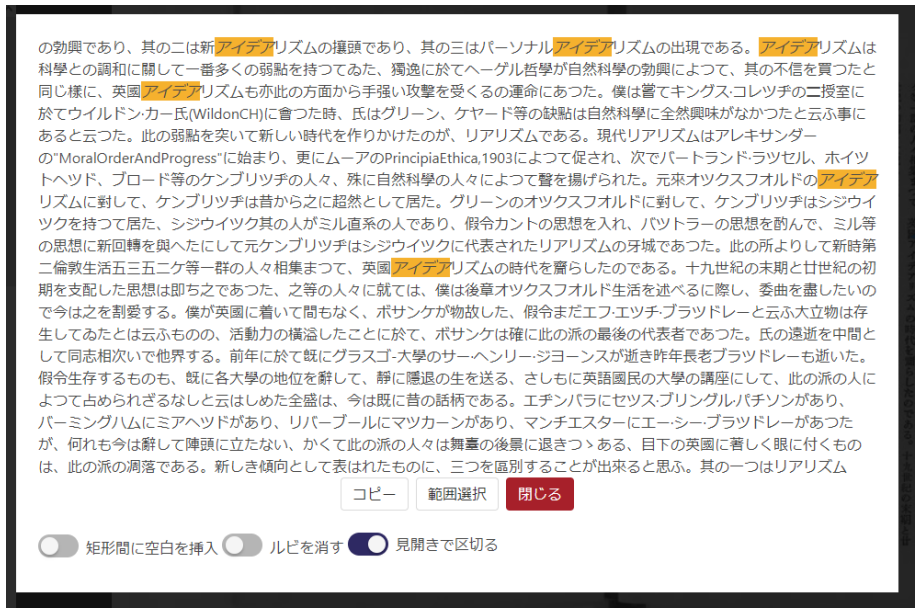
https://lab.ndl.go.jp/dl/book/1437246?keyword=クリエイティブ&page=42

OCRテキストデータのダウンロードボタン

【サービス①】
次世代デジタルライブラリー

- OCRテキスト表示機能・コピペ機能





https://lab.ndl.go.jp/dl/book/980819?keyword=アイデア&page=36

【サービス①】
次世代デジタルライブラリー

NDLラボで開発した機械学習モデルにより、資料中の図版について、類似図版の検索も可能

https://lab.ndl.go.jp/dl/illust/search?image=2535667_5_1

【サービス②】
NDL Ngram Viewer

https://lab.ndl.go.jp/ngramviewer/

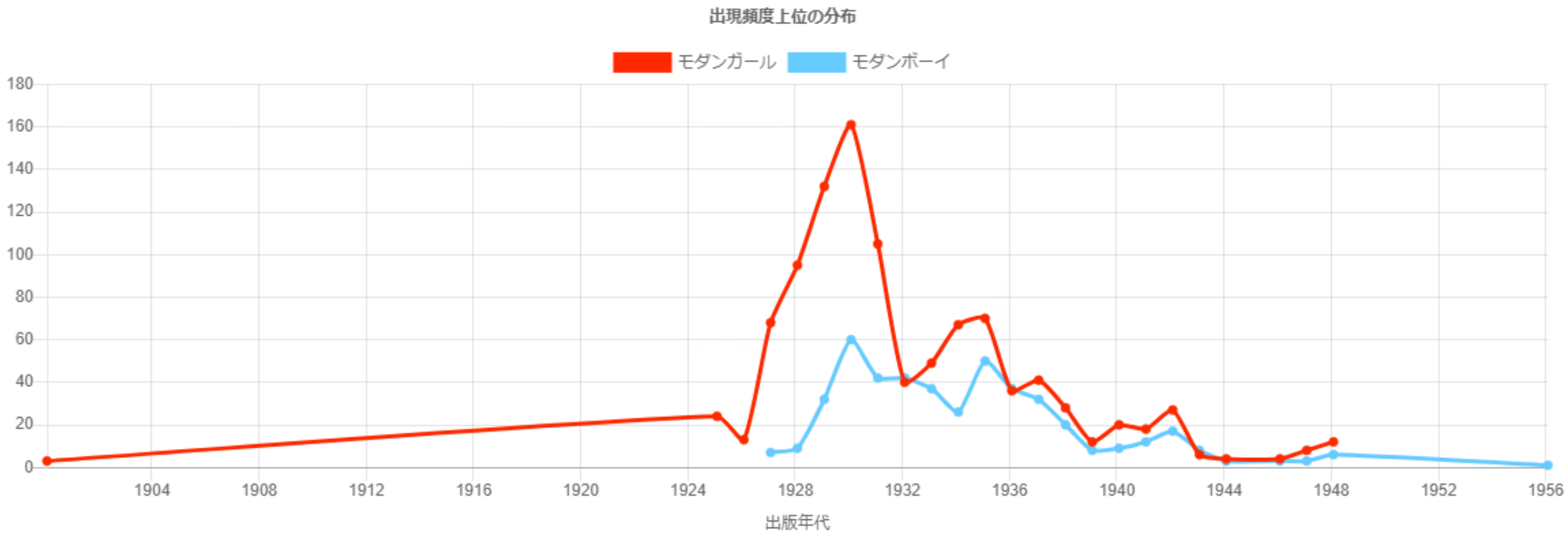
国立国会図書館デジタルコレクションでインターネット公開している資料のうち、著作権保護期間が満了した図書資料約28万点（次世代デジタルライブラリーの全文検索対象と同じ資料群）のOCRテキストデータから集計した、約8.3億種類の単語及びフレーズの出現頻度を、出版年代ごとに可視化できるツール

何ができるサービス？

- **OCRテキストデータ中に存在する単語や語用法の出現頻度の可視化**
- **正規表現による周辺キーワードの探索**

【サービス②】
NDL Ngram Viewer

「モダンボーイ」と「モダンガール」の出現頻度を比較



https://lab.ndl.go.jp/ngramviewer/?keyword=%E3%83%A2%E3%83%80%E3%83%B3%E3%82%AC%E3%83%BC%E3%83%AB%2F%E3%83%A2%E3%83%80%E3%83%B3%E3%83%9C%E3%83%BC%E3%82%A4&size=100&from=0

【サービス②】
NDL Ngram Viewer

例えば「関ケ原の合戦」について調べたいとき

「関ケ原」？
「関ヶ原」？
それとも「関が原」？

「関ヶ原の合戦」？
「関ケ原合戦」？
それともひょっとして
「関ケ原の戦い」？

【サービス②】
NDL Ngram Viewer

→「関.原.*戦い?」で正規表現検索！

https://lab.ndl.go.jp/ngramviewer/?keyword=%E9%96%A2.%E5%8E%9F.%2a%E6%88%A6%E3%81%84%3F&size=100&from=0

キーワード	総出現頻度	
関ケ原の戦	2010	次世代デジタルライブラリーで検索
関ケ原合戦	1258	次世代デジタルライブラリーで検索
関ヶ原の戦	967	次世代デジタルライブラリーで検索
関ヶ原合戦	824	次世代デジタルライブラリーで検索
関ケ原の一戦	626	次世代デジタルライブラリーで検索
関ケ原の合戦	466	次世代デジタルライブラリーで検索
関ケ原戦	431	次世代デジタルライブラリーで検索
関が原の戦	394	次世代デジタルライブラリーで検索
関ケ原ノ戦	352	次世代デジタルライブラリーで検索
関ケ原大戦	346	次世代デジタルライブラリーで検索
関ヶ原の合戦	284	次世代デジタルライブラリーで検索
関ケ原の大戦	201	次世代デジタルライブラリーで検索
関ヶ原戦	181	次世代デジタルライブラリーで検索

【サービス②】
NDL Ngram Viewer

演算子	クエリ例	説明
.	修.者	任意の1文字を表す
*	日*精進	直前の表現が0個以上あることを表す
+	郵便局+長	直前の表現が1個以上あることを表す
?	巡査部?長	直前の表現が0個か1個あることを表す
{,}	私を.{2,3}にする	直前の文字の繰り返し回数の範囲を指定する。例1:{2,4}:2回以上4回以下 例2:{3,}:3回以上 例3:{,5}:5回以下
[…]	大正[ーア-ン]{6,}	ブラケット内の1文字を表す。「-」で範囲を、「^」で否定を表す。例1[abc]:a,b,cのうち1文字 例2[ア-ン]：アからンまでの1文字（カタカナいずれか1文字） 例3[^ア-ン]：アからンまで以外の1文字（つまりカタカナ以外の1文字）
(…)	春の海ひねもす(のたり)*	かっこで囲んだ範囲のグループを形成し、単一の文字として扱う。（他の演算子と組み合わせて用いる）
|	ご(機嫌|きげん)よう	左辺または右辺の最長のパターンにマッチすることを表す

Q.著作権保護期間の存続している資料のデータは使えないの？

A.今日のところは残念ながら提供できませんが、条件を満たせばお渡しできます！

- 著作権保護期間の存続している資料から作成したOCR学習用データセットについては、原資料の著作権保護の観点からGitHub上に公開することはできません。

- これらのデータに関しては、当館との協議のうえで著作権法上認められた範囲内での利用（著作権法第30条の4の規定による機械学習目的など）に限り、当館と書面を取り交わした上で提供することが可能です。

Q. OCRテキストデータやNDLOCRの性能はどのくらい？

A.事業の報告資料を公開していますのでご覧ください！

OCRテキストデータの報告資料
https://lab.ndl.go.jp/data_set/ocr/r3_line/

NDLOCRの報告資料
https://lab.ndl.go.jp/data_set/ocr/r3_morpho/

Q.提供されているデータセットを利用して商用サービスを開発しても大丈夫？

A. 問題ありません、大丈夫です！

データセットのライセンスはいずれもCC BY 4.0で提供しています

NDLOCRについても、新規開発部分はCC BY 4.0、既存のライブラリ等を利用している部分については寛容型オープンライセンスのものを採用しているため、商用非商用を問わず自由な改変、利用が可能です。

アイデアソン参加者の皆さまへ

- 同じ班の誰かが良い知恵を持っていらっしゃるかもしれません、是非、「こんなことできたら嬉しいな」を気軽に投げかけてください（特に前半のアイデア出しの段階）

- 専門知識をお持ちの方は、「こうしたらうまくいくよ・こんな事例があるよ」を積極的に共有して頂けると大変ありがたいです

- 今後データセットを使って共同研究等を進めていく仲間づくりの場としていただいても大歓迎です

アイデア出しのコツ

・プレイズ・ファースト（Praise First）

「先に褒めよ」

アイデアの良い所を褒める。批判はあと

→アイデアを育てる

「プレイズ・ファースト」の本質は「良点発見」（ブレインストーミングのルール＋α）.
石井力重の活動報告.http://ishiirikie.jpn.org/article/67402619.html (2022-06-08 accessed)

一般参加者の皆さまへ

- ブレイクアウトセッション中は、ぜひチャット機能を活用して活発に質問やコメントをしてください

- 質疑の時間では、アイデアの実現につながるようなご意見を歓迎します

- もしよろしければ、SNS等にも書き込んで頂ければと思います

  ハッシュタグ： **#JOSS2022 #90分アイデアソン**

全ての参加者の皆さまへ

- 今回のアイデアソンによって、集まったいろいろなバックグラウンドの皆さんの間で交流が生まれ、オープンデータから新しいサービスが育っていくきっかけになることを願っています

- 一人では思いつかない素敵なアイデアを議論の中に発見して頂ければと思います

- アイデアソンが終わった後が本番……になると嬉しいです。イベント終了後も気兼ねなく、是非是非お問い合わせください

さいごに

アンケートにご協力お願いします
https://enquete.ndl.go.jp/249638

お問い合わせは、NDLラボ lab@ndl.go.jp まで

NDL Lab
National Diet Library, Japan